Enhancing global positioning by image recognition

David Marimon* Tomasz Adamek† Arturo Bonnin Tomasz Trzcinski

Telefonica Research and Development
Barcelona, Spain

ABSTRACT

Current commercial outdoor Mobile AR applications rely mostly on GPS antennas, digital compasses and accelerometers. Due to imprecise readings, the 2D placement of points of interest (POI) on the display can be uncorrelated with reality.

We present a novel method to geo-locate a mobile device by recognizing what is captured by its camera. A visual recognition algorithm in the cloud is used to identify geo-located reference images that match the camera's view. Upon correct identification, fusion of sensor data from the mobile device and the server is used to establish the most probable location. The corrected geo-location can then be used on the mobile device for further POI representation or as a better initialization step for further continuos tracking.

Index Terms: I.5.4 [Pattern Recognition]: Applications—Computer vision; I.4.8 [Image Processing and Computer Vision]: Scene Analysis—Object recognition and Sensor fusion;

1 INTRODUCTION

During the past three years there has been an explosion of commercial outdoor Mobile Augmented Reality (MAR) applications that commonly depend on GPS antennas, digital compasses and accelerometers. These sensors provide the geo-location of the mobile user and the direction towards which the camera of the device is pointing. This direction is enough to show geo-located points of interest (POIs) on the mobile display overlaid to the video feed from the camera.

Due to non-accurate readings, the 2D placement of POIs on the display can be uncorrelated with reality. This is especially dramatic for POIs that are close to the user. We could easily imagine the situation where a GPS provides a location that is on the other side of the corner in a POI-crowded area. The display would not provide information about the POI that is just in front of the user. Such situation can impoverish not only user experience but also hyper-local Mobile AR services.

Recent research on outdoor augmented reality has mostly focused on visually recognizing and registering pose to natural features in the scene [9, 8, 11, 2]. Although highly accurate 6DOF pose estimation can be achieved, those techniques rely on available sets of images of those landmarks that are being augmented.

However, the current situation of MAR applications is slightly different. As a matter of fact, most displayed POIs come from content providers with data sets that do not usually have a reference image associated (or at least not necessarily one of its outside facade).

Fortunately, there exist data sets of images that are geo-referenced (e.g. Panoramio[1] or Google Street View[2]). Therefore, we propose to exploit this sort of data sets to enhance GPS-based geo-location with image recognition.

In particular, our contributions can be summarized as follows:

- A novel mechanism to fuse data provided by GPS, compass and accelerometers with single or multiple recognition of geo-referenced images.
- A client-server architecture intended for initialization (and re-initialization after loss of track) for systems that perform continuous tracking. Such architecture would enable correct 2D positioning of POIs on the AR view of a mobile display. In fact, the improved geo-localiastion can even benefit visualization of places without a reference image associated.

The structure of the paper is as follows. Next section describes prior art on outdoor localization based on visual features. Section 3 describes the proposed method. First, an overview of the architecture is presented. Second, the visual recognition engine is described. Third, the fusion of data from different sensors is explained. Section 4 describes the validation of the proposed method. Final remarks and future paths of research are provided at the end.

2 RELATED WORK

Early examples of outdoor mobile augmented reality used data from non-visual sensors [3].

Recent advances in computer vision have enabled online tracking of natural features for outdoor augmented reality [9, 2].

Reitmayr and Drummond [9] presented an edge-based approach to track street facades based on a rough 3D model. This approach was further enhanced with an initialisation mechanism based on an accurate GPS antenna [8].

More recently, Arth *et al.* [2] presented a 6DOF tracking algorithm that performs wide area localization based on Potentially Visible Sets of 3D sparse reconstructions of the environment. The system runs on a mobile device and counts on external initialization. For outdoors, the authors propose to employ GPS.

The methods cited above are focused on precise online tracking where reference features are available on the device prior to start tracking. Another path to offer augmentation of the video feed is by recognizing landmarks in front of the camera. Instead of online tracking and registering, pose is computed by detection.

In this regard, Schindler *et al.* [10] presented a recognition method for large collections of geo-referenced images. The method builds on vocabulary trees of SIFT features [4] and inverted file scoring as in [6].

Takacs *et al.* [11] present a system that performs keypoint-based image matching on a mobile device. In order to constrain the matching, the system quantizes the user's location and only considers nearby data. Features are cached based on GPS and made available for online identification of landmarks. Information associated to the top ranked reference image is displayed on the device.

3 GEO-LOCATING BY IMAGE RECOGNITION

We propose an initialisation mechanism for MAR applications based on geo-referenced image recognition. The system is a client-server framework where all the computation is performed on the server side (see Figure 1).

*e-mail: marimon@tid.es
†e-mail: tomasz@tid.es

[1] www.Panoramio.com
[2] www.google.com/streetview

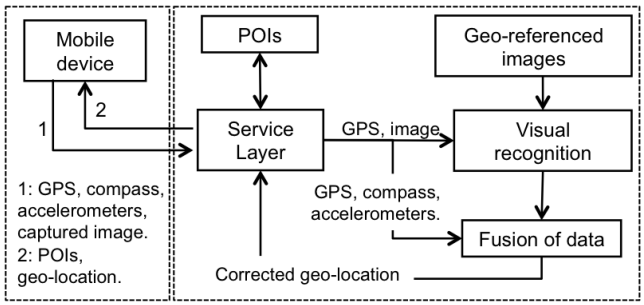

Figure 1: Block diagram of the client-server framework. All computation occurs on the server side.

Although there are good reasons in MAR for balancing the computation towards the mobile device (such as scalability and latency), bear in mind that our method is designed for initial localization. Therefore, little bandwidth is consumed (circa 50-75 KB) and delay during this phase is not so critical for the user. As a counterpart, with our architecture, we gain database flexibility and can perform more complex visual recognition tasks regardless of the mobile computing power.

We also underline the fact that the proposed method is complementary to the approaches described in the previous section. On one hand, our approach could be used for initialization of those online tracking algorithms running on mobile phones where real-time registration is key for the AR experience. On the other hand, our system can not only display the POIs that are image-tagged (as in [11]), but also benefit augmentation of those POIs that do not have a reference image because of the improved positioning that our system can provide.

3.1 Visual recognition of geo-referenced images

The visual recognition engine is the core technology that permits to identify all similar images and their spatial relation with respect to the image captured by the mobile device. Of course, taking the advantage of the GPS information, the recognition is constrained to reference images that were captured close to the query image.

Our approach, similarly to other state of the art solutions, relies on local visual features (DART keypoints in our case [5]), hierarchical dictionaries of visual words [6], inverted file structures, and a spatial verification stage of the top ranked initial results [7].

However, in contrast to the above-mentioned methods, the commonly used TF-IDF scoring mechanism [7] is replaced by an early clustering of matches in the pose space (limited to orientation and scale), in a way similar to the voting in the well known Hough transform [1].

More precisely, the matching is performed in two stages: (i) initial ranking of reference images using visual words and voting in the limited pose space, and (ii) re-ranking of the initial results using more complex spatial verification stage similar to the one proposed in [7].

We have found that the inclusion of even such a rudimentary spatial verification mechanism in the very initial stage of the recognition is helpful in cases of small objects buried within complex scenes. This engine has been evaluated for the purpose of near-duplicate video copy detection and mobile visual search [1].

3.2 Fusion of sensors and visual recognition

The module that fuses data is responsible for obtaining the corrected longitude and latitude coordinates. The proposed method projects all sensor data into references w.r.t.a 2D map of longitude and latitude coordinates.

For each reference image that matches the query one, we can obtain a geometric spatial relation in the form of a similarity transformation[3].

The similarity transformation provides one aspect that is relevant for our system: scale (λ). Scale is used here to determine how close the user is to a location where a reference image in the database was taken. Since scale cannot be translated to GPS coordinates, we transform it into a measure of belief.

Translation, on the other hand, is of little use for us since a simple camera panning motion could be confused with user's displacement. Therefore, we do not transform that into a change in geo-coordinates. For rotation, a similar rationale is followed.

The compass and accelerometers are used to determine the direction of sight onto the 2D map. This direction provides further belief on scale changes depending on the coordinates $\mathbf{i}$ of a matched image and those provided by the GPS antenna $\mathbf{s}$.

The process consists in the following steps:

1. establish the vector $\vec{v}$ from $\mathbf{s}$ to $\mathbf{i}$;
2. establish the angle θ between the direction of sight and $\vec{v}$;
3. determine the influence of $\mathbf{s}$ and $\mathbf{i}$ depending on angle and scale; and
4. consider multi-recognition influences.

The influence mentioned in step 3 is the influence factor n computed as follows:

$$n = \begin{cases} \sqrt{w}/K & \text{if } \theta \in [-\pi/4, \pi/4] \text{ and } \lambda \geq 1 \\ & \text{or if } \theta \in [3\pi/4, 5\pi/4] \text{ and } \lambda \leq 1 \\ w/K & \text{otherwise.} \end{cases} \quad (1)$$

where $w = e^{-(\lambda-1)^2/\sigma^2}$ for $\lambda \in [0,2]$ and $w = 0$ otherwise, σ is chosen experimentally maintiaining a narrow bell shape in w, and K is the number of top-ranked reference images considered. K depends on scored recognition level. This influence factor n permits to limit the contribution of recognition to image matchings that have similar scale and therefore where probably taken from a place close to that of the query.

Corrected coordinates are obtained considering all n_k influences together with GPS:

$$(\text{long,lat}) = \sum_k (n_k \cdot \mathbf{i}_k + (K^{-1} - n_k) \cdot \mathbf{s}). \quad (2)$$

Note that only images that are physically close to the query are considered (see previous section). Therefore, we can assume that only pictures taken with similar focal length contribute positively (similar scale) to correct the coordinates.

4 Experiments

In order to validate the proposed technique, we have conducted two experiments. First, we have evaluated the performance of the visual recognition engine with a database of outdoor environment pictures. Second, we have experienced with geo-positioned data acquired from public repositories and test query images from a mobile device.

4.1 Visual recognition of outdoor environments

For the purpose of evaluating the performance of the visual recognition module within an outdoor environment we have used The Oxford Buildings Dataset [12]. This dataset consists of five thousand reference images of building facades together with ground truth data.

[3] Although epipolar geometry could be used, we prefer to constrain the system to the realistic situation that camera calibration is not available for any of the managed images.

Figure 2: Top ranked images (left-to-right and top-to-bottom) retrieved from the Oxford dataset in response to a query region (top-left image).

For this database, the initial pose voting stage alone (see Section 3.1) obtains recognition results that are similar to other methods using more complex spatial verification stages (e.g. [7]). In particular, we obtain a MAP of 0.47. At the same time, it should be stressed that computational cost of our voting scheme compares favorably to the TF-IDF scheme. An example of the top-ranked results is shown in Figure 2.

4.2 Enhancing global positioning

This experiment consists of a qualitative evaluation of the correction of global positioning performed by the proposed mobile client-server system. The goal is to observe how much improvement is possible over data provided by a GPS antenna on a mobile device, when using our system fusing GPS data and a query picture taken by this same mobile device.

In order to validate our system, we must first acquire a database of images that is reliable enough. We have experimented with two datasets: Panoramio and Google Street View. Both have available APIs to retrieve images around certain locations. We have picked several locations around well-known touristic places in the city of Barcelona, Spain.

The other side of the validation resides on the client side and the capacity to generate query data to test our system. For that, we have used an *ad hoc* application that runs on a Nexus One mobile phone. This application shoots pictures and stores them together with the corresponding GPS readings of the mobile device.

The reference dataset obtained from Panoramio turned out to be unreliable. Panoramio has user-generated geo-location and that is not accurate enough for our purposes. The one obtained from Street View is much more accurate and served for the purpose of correction.

In the following, we explain the different scenarios that we observed. The term *reasonable GPS data* refers to the situation were data was acquired in wide streets and therefore we assume that several satellites were visible. Unreliable data refers to the opposite case.

- Distinctive facade; reasonable GPS data. In this case, we observe that the fusion is beneficial. Having strong evidence of visual recognition permits to move towards readings performed by more sophisticated GPS equipment such as that used for Street View. The correction observed can be up to several meters. Figure 3 shows an example of this situation. The red R symbolizes a location with an associated reference image that was not considered a match. The green R is a matched reference image (in this example there is only one but could be multiple). The yellow Q is the GPS reading of the mobile device. The blue Q is the result of our fusion system. Also marked is the location of the manually obtained ground truth and the query picture considered.

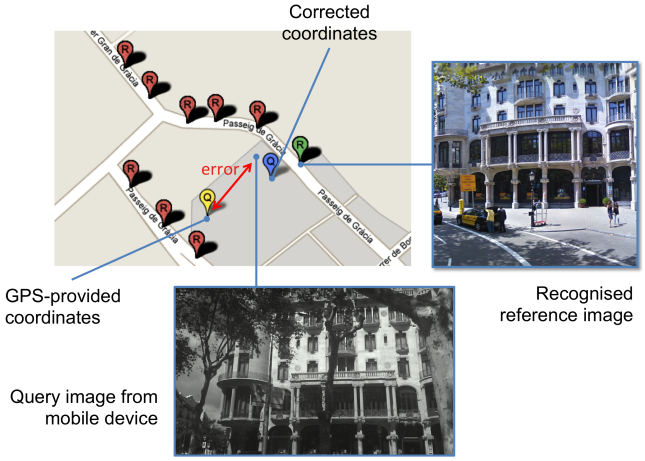

Figure 3: Example correction of global positioning through visual recognition of geo-referenced images. See Section 4.2 for a complete description of the figure.

- Repetitive facade or environment; reasonable GPS data. In this case, elements such as trees or simply the shape of the facade do not provide consistent visual evidence and GPS data alone is as good as the result of our fusion. Note that in case of low visual recognition scores our fusion method tends to favor GPS data. This ensures that the fusion results are not worse than the initial GPS information available on the mobile device.

- Unreliable GPS data. In small streets, the mobile GPS antenna has more difficulties to provide accurate readings. In this case, we have seen a significant improvement with our system, especially for distinctive facades.

This experiment has served our purpose of evaluating the scenarios where our fusion system is more beneficial and how it can improve initialization of more sophisticated tracking systems.

It must be pointed out that some limitations of the reference dataset impoverished the recognition task. More specifically, reference images from Street View are portions of 360-degrees images. Recognition of many facades was affected by the strong deformation of above-street level parts of buildings.

5 Conclusions

We have presented a client-server architecture and a method to correct mobile GPS readings with image recognition. A server-side visual recognition engine enables the fusion of mobile sensors with multiple image matching.

We have evaluated the performance of visual recognition of outdoor environments with successful results. We have also confronted the whole system to real-world data acquired from public datasets and tested with an application running on a mobile phone. Our experiments show that our system delivers different improvement levels depending on the scenario. Overall, our system can correct GPS readings when enough visual evidence is available.

With those experiments, we have addressed our purpose of demonstrating on one hand that the proposed fusion is able to assist tracking initialization in general and therefore complement other state-of-the-art solutions. On the other hand, from an application perspective, our system can be used for applications augmenting POIs that do not necessarily have reference images associated; and the system is also interesting for the correct augmentation of POIs that are near the user.

Future paths of research will focus on quantitatively measuring the metric improvement of our system w.r.t. GPS data provided by off-the-shelft mobile devices.

Acknowledgements

The authors would like to thank Roberto Álvarez from Indra Software Labs for developing the mobile application.

This work was developed within the MobiAR project financed by the MITyC inside the Avanza program.

Telefónica I+D participates in Torres Quevedo subprogram (MICINN), co-financed by the European Social Fund, for Researchers recruitment.

References

[1] T. Adamek and D. Marimon. Large-scale visual search based on voting in reduced pose space with application to mobile search and video collections. In *Proc. Industrial Program of the 2011 IEEE Intl. Conf. on Multimedia & Expo (ICME 2011)*, 2011.

[2] C. Arth, D. Wagner, M. Klopschitz, A. Irschara, and D. Schmalstieg. Wide area localization on mobile phones. In *Proc. Intl. Symp. on Mixed and Augmented Reality (ISMAR)*, pages 73–82, 2009.

[3] R. Azuma, B. Hoff, H. Neely, and R. Sarfaty. A motion-stabilized outdoor augmented reality system. In *Proc. IEEE Virtual Reality (VR)*, pages 252–259, 1999.

[4] D. Lowe. Distinctive image features from scale-invariant keypoints. *Intl. Journal of Computer Vision*, 60(2):91–110, 2004.

[5] D. Marimon, A. Bonnin, T. Adamek, and R. Gimeno. DARTs: Efficient scale-space extraction of DAISY keypoints. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, pages 2416–2423, San Francisco, June 2010.

[6] D. Nister and H. Stewenius. Scalable recognition with a vocabulary tree. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, pages 2161–2168, 2006.

[7] J. Philbin, O. Chum, M. Isard, J. Sivic, and A. Zisserman. Object retrieval with large vocabularies and fast spatial matching. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2007.

[8] G. Reitmayr and T. Drummond. Initialisation for visual tracking in urban environments. In *Proc. Intl. Symp. on Mixed and Augmented Reality (ISMAR)*, pages 161–160, 2007.

[9] G. Reitmayr and T. W. Drummond. Going out: Robust tracking for outdoor augmented reality. In *Proc. Intl. Symp. on Mixed and Augmented Reality (ISMAR)*, pages 109–118, 2006.

[10] G. Schindler, M. Brown, and R. Szeliski. City-scale location recognition. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, volume 0, pages 1–7, 2007.

[11] G. Takacs, V. Chandrasekhar, N. Gelfand, Y. Xiong, W.-C. Chen, T. Bismpigiannis, R. Grzeszczuk, K. Pulli, and B. Girod. Outdoors augmented reality on mobile phone using loxel-based visual feature organization. In *Proc. Multimedia Information Retrieval*, pages 427–434, 2008.

[12] Visual Geometry Group. The Oxford Buildings Dataset. `http://www.robots.ox.ac.uk/~vgg/data/oxbuildings/`.